Published online 04 December 2014 *Nucleic Acids Research, 2015, Vol. 43, Database issue* ***D1163–D1170***
doi: 10.1093/nar/gku1244

BioAssay Research Database (BARD): chemical biology and probe-development enabled by structured metadata and result types

E.A. Howe[1], A. de Souza[1,*], D.L. Lahr[1], S. Chatwin[1], P. Montgomery[1], B.R. Alexander[1], D.-T. Nguyen[2], Y. Cruz[3], D.A. Stonich[4], G. Walzer[1], J.T. Rose[1], S.C. Picard[1], Z. Liu[1], J.N. Rose[1], X. Xiang[1], J. Asiedu[1], D. Durkin[1], J. Levine[1], J.J. Yang[5], S.C. Schürer[6], J.C. Braisted[2], N. Southall[2], M.R. Southern[3], T.D.Y. Chung[4], S. Brudz[1], C. Tanega[2], S.L. Schreiber[1], J.A. Bittker[1], R. Guha[2,*] and P.A. Clemons[1,*]

[1]Center for the Science of Therapeutics, Broad Institute, 415 Main Street, Cambridge, MA 02142, USA, [2]National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH), 9800 Medical Center Drive, Rockville, MD 20850, USA, [3]The Translational Research Institute, The Scripps Research Institute, 130 Scripps Way, Jupiter, FL 33458, USA, [4] *Conrad Prebys* Center for Chemical Genomics, Sanford|Burnham Medical Research Institute, 10901 N. Torrey Pines Road, La Jolla, CA 92037, USA, [5]University of New Mexico Center for Molecular Discovery, University of New Mexico Health Sciences Center, 2500 Marble Avenue NE, Albuquerque, NM 87131, USA and [6]Center for Computational Science, University of Miami, 1320 S. Dixie Highway, Gables One Tower, Coral Gables, FL 33146, USA

Received August 29, 2014; Revised November 11, 2014; Accepted November 12, 2014

ABSTRACT

BARD, the BioAssay Research Database (https://bard.nih.gov/) is a public database and suite of tools developed to provide access to bioassay data produced by the NIH Molecular Libraries Program (MLP). Data from 631 MLP projects were migrated to a new structured vocabulary designed to capture bioassay data in a formalized manner, with particular emphasis placed on the description of assay protocols. New data can be submitted to BARD with a user-friendly set of tools that assist in the creation of appropriately formatted datasets and assay definitions. Data published through the BARD application program interface (API) can be accessed by researchers using web-based query tools or a desktop client. Third-party developers wishing to create new tools can use the API to produce stand-alone tools or new plug-ins that can be integrated into BARD. The entire BARD suite of tools therefore supports three classes of researcher: those who wish to publish data, those who wish to mine data for testable hypotheses, and those in the developer community who wish to build tools that leverage this carefully curated chemical biology resource.

INTRODUCTION

High-throughput screening enables rapid measurement of the effects of large numbers of small molecules on biochemical or cell-based biological systems. Such experiments result in large datasets, containing thousands to millions of measurements, which enable identification of candidate small molecules for further development as tool compounds or drug leads. Several bioassay data-storage and data-analysis systems exist that serve the chemical biology, computational chemistry, and bioinformatics communities. The National Cancer Institute and Harvard Institute of Chemistry and Cell Biology partnered to develop ChemBank, a first-generation chemical biology data resource (1,2). ChemBank was built to house screening data and to provide a cheminformatics platform for analysis of those data. The National Institutes of Health (NIH) and the National Center for Biotechnology Information (NCBI) partnered to develop PubChem to provide a national deposition repository for related data emanating from the Molecular Libraries Program (MLP) (3,4), and to create tools enabling its use as a data-analysis portal. Recently, substantive updates have been made to the PubChem datasets and infrastructure, es-

*To whom correspondence should be addressed. Tel: +1 617 714 7346; Fax: +1 714 8969; Email: pclemons@broadinstitute.org
Correspondence may also be addressed to A. de Souza. Tel: +1 617 714 7000; Fax: +1 617 714 8102; Email: Andrea.desouza@sloan.mit.edu
Correspondence may also be addressed to R. Guha. Tel: +1 301 217 5733; Fax: +1 301 827 2534; Email: guhar@mail.nih.gov

pecially with the introduction of panel assays and a new categorized comment system that allows optional annotation of assays to comply with domain-specific data standards (5). The ChEMBL database (6,7) demonstrated the utility of annotating data consistently across experiments and cross-referencing compound targets against external databases, such as Gene Ontology (GO, (8)) and the Kyoto Encyclopedia of Genes and Genomes (KEGG, (9)).

Despite these early successes in chemical biology data sharing, the management and reuse of these data has remained difficult due to a lack of standardization, particularly of assay descriptions, which prevents effective, structured query across datasets. While researchers reading the descriptive text of an analysis may not be hindered by such inconsistent annotations in different assays, cross-assay computational analysis of datasets with inconsistent (or non-existent) meta-data is impossible. The lack of standardized metadata annotation therefore restricts analyses of the data to singleton datasets and prevents the systematic, unbiased survey of the entire contents of a database. Datasets such as those produced by the MLP are of high quality and scientifically valuable, but were not annotated or formatted consistently across screening centers despite a stated goal of MLP to analyze the resulting data collectively.

In this report, we present BioAssay Research Database (BARD), a freely available system for registering and querying publicly accessible bioassay data. BARD was commissioned by the MLP to be the principal analysis environment housing screening data generated by its probe-development projects. It currently holds complete screening data for over 600 projects from program screening centers. These projects consist of primary high-throughput screens, confirmatory assays, and counter-screens testing on average over 150,000 compounds. Each project has been curated to conform to a flexible, consistent vocabulary that includes a standard representation and common language for organizing bioassays and their results. This curation has enabled systematic, cross-assay analysis of datasets regardless of the identity of the contributing center or individual data submitter. While other bioassay databases annotate their data with some shared terms, BARD uses a controlled vocabulary specifically developed to describe assay protocols. The data housed within PubChem lacks the consistent annotation required for effective search across datasets, especially analyses designed to reveal methodological connections between unrelated projects. Similarly, ChEMBL lacks extensive structured annotation of assays and assay protocols, making assembly of data from similar experiments difficult. These unstructured data are not well-suited to computational search and analysis across assays and projects. BARD is a tool that researchers can use to share chemical biology data and develop hypotheses on the influence of chemical probes on biological systems, suggesting future experiments to pursue.

RESULTS

BARD supports three distinct research audiences: assay-data depositors, data miners, and software developers (Figure 1). Depositors can publish their bioassay data using a guided data-entry tool that assists them, ensuring that new

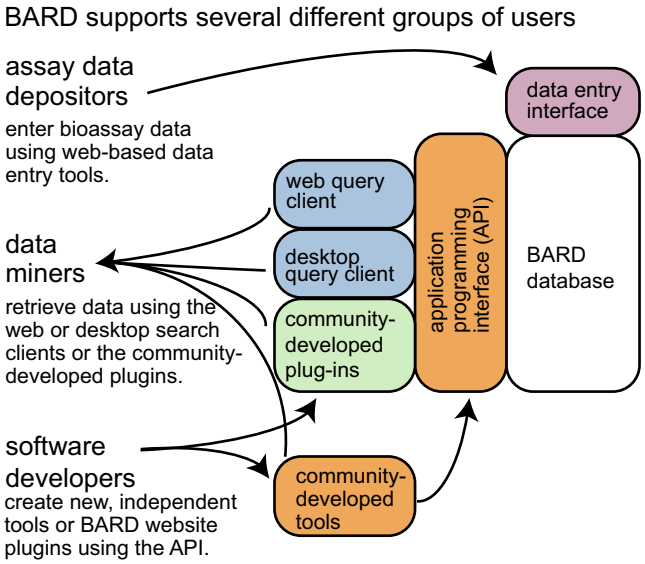


Figure 1. BARD data access and entry. BARD data are entered by data depositors, using the web-based data-entry interface (purple). Data miners can explore BARD data using several tools, developed both by BARD developers and the larger developer community (blue and green). Software developers can create new tools that access BARD data using the API and deploy them either as stand-alone tools or as plug-ins that can be integrated into the BARD website (green and orange).

records are well-structured and complete. Researchers can then mine these newly deposited data as part of the full BARD database, using a powerful search interface and set of visualization and statistical tools. Developers can build new research tools either as standalone programs, or as plug-ins integrated into BARD using a well-documented application-programming interface (API).

Data migration and structure

We curated and migrated 630 MLP datasets from PubChem into BARD. This data-migration process involved converting project metadata to the BARD structured vocabulary and ensuring a consistent data structure across all assays. Though we built tools to partially automate this process, manual intervention was required to successfully import most assay protocols. Curators interpreted free-text PubChem entries and refactored those metadata using the BARD vocabulary. Following initial migration using automated scripts, annotations and changes were reviewed by MLP scientists involved in the original experiments (Table 1). In addition, long-form free-text was migrated along with the curated meta-data. This migration and curation process served as the template for the data-entry tools that are now built into BARD.

The BARD vocabulary was built to address the problem of inconsistent, incomparable annotations between datasets. Complex queries of bioassay data require a common vocabulary and consistent data structures across all assays. Because such structures were previously unavailable or under-developed, we created an extensive formal vocabulary and term hierarchy for BARD (10). For example, because PubChem allows capture of context items in free-text

Table 1. BARD datasets

Center	Projects	Experiments	Assays	Average compounds per project
NIH Chemical Genomics Center – Molecular Libraries Probe Production Center	231	915	579	204 927
The Scripps Research Institute – Molecular Screening Center	85	1922	1341	232 921
The Broad Institute Therapeutics Platform	61	1166	702	284 454
Sanford-Burnham Medical Research Institute – Conrad Prebys Center for Chemical Genomics	91	676	480	233 728
University of New Mexico Center for Molecular Discovery	44	530	289	129 141
Southern Research Molecular Libraries Screening Center	33	260	101	116 317
Johns Hopkins University Drug Discovery	22	260	128	307 537
Vanderbilt Institute of Chemical Biology	21	371	331	51 863
University of Pittsburgh Molecular Libraries Screening Center	10	46	21	125 871
The Penn Center for Molecular Discovery	16	47	30	103 172
Emory Chemical Biology Discovery Center	16	37	26	102 631
Total	**630**	**6230**	**4028**	**13 011 151**[a]

Summary of datasets migrated into the BARD database, grouped by contributing screening center.
[a]Sum of all compounds in all projects.

fields, the concept of "percent inhibition at 10 μM" is represented semantically 1,800 different ways in the various MLP depositions in that database. Formulating a search function that can accommodate all of these variants in a single query is difficult. After curation to unify these annotations under the BARD vocabulary, however, a search could include the criterion of "percent inhibition at 10 μM" and reasonably expect to retrieve all assays that included any one of the 1,800 semantic variations of the same concept. This standardization of metadata terms was done across all of the metadata in the 630 MLP datasets.

The BARD vocabulary ensures preservation of the meaning of assay definitions, while also providing flexibility, extensibility, and maintainability. The vocabulary is modifiable and enforces consistent naming of annotations, such as for cell lines, avoiding some limitations inherent to free-text data. The vocabulary alleviates the problem of different research groups using multiple different names for the same concept or entity (https://bard.nih.gov/BARD/element/showTopLevelHierarchyHelp). The BARD vocabulary leverages and references established external ontologies wherever possible, including BioAssay Ontology (BAO) (11,12), GO, and KEGG. It includes dedicated data structures for information like incubation times, compound concentrations, and assay types. Further, storage of these data in dedicated fields means that queries can require that a search term belong to a particular data type, rather than the merely requiring the existence of a term in any free-text descriptive field.

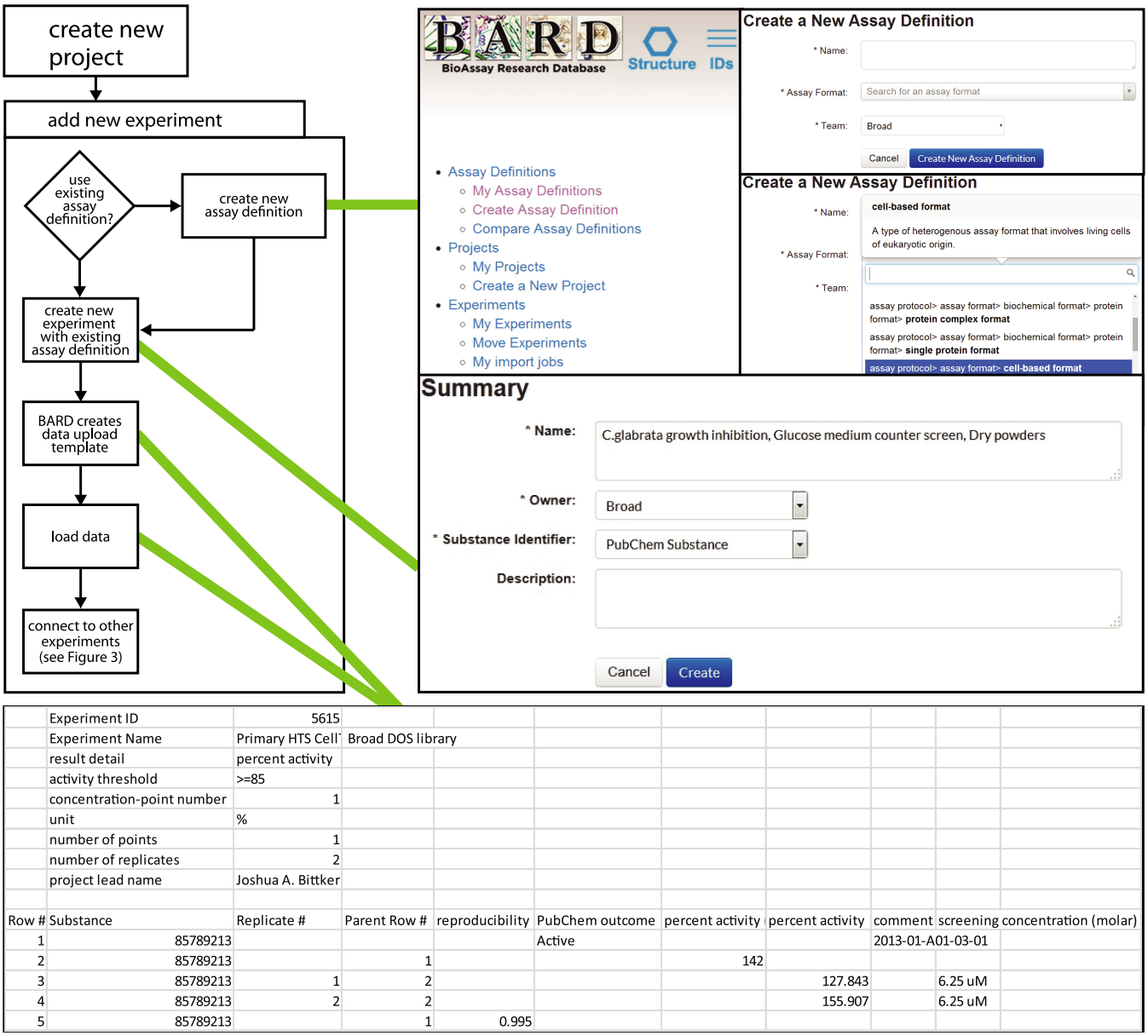
Technology

BARD comprises several independent modules integrated by APIs. The data-entry component of BARD is integrated into the BARD website, and built with the Grails web-development framework (https://grails.org/). Layout support was implemented using Twitter Bootstrap (http://getbootstrap.com/2.3.2/). Data entered into BARD are first checked for data consistency using business rules and then stored in an Oracle database. An extract-transform-load (ETL) process transfers the curated data from this database to a data warehouse, which stores the data in a form optimized for fast query. The warehouse is a MySQL (http://www.mysql.com/) database supported by Solr (http://lucene.apache.org/solr/) for fast search. Fast chemical structure search in BARD is supported by a custom, in-memory solution that can scale to very large databases (https://tripod.nih.gov/?p=361). Local copies of the GO and KEGG databases are also stored, for fast access by auto-complete functions.

A representational state transfer (REST) API (https://github.com/ncats/bard/wiki) mediates all transactions with the warehouse database and returns JSON-formatted data. This API is the primary access point for the data warehouse and supports queries from the Web Query tool and the Desktop Client. The API also supports a plug-in system, allowing external developers to create new functionality that can be added to the BARD core API. It is implemented using Java servlets and provides a REST interface to the BARD data warehouse. The REST interface employs the Jersey (https://jersey.java.net/) library to provide the requisite scaffolding and routing mechanisms. Jackson (http://jackson.codehaus.org/) is employed to support processing of JSON documents, which is the primary response format for the bulk of the API.

The browser-based query client hosts a wide variety of query and data-mining functions for browsing and searching the data stored in the data warehouse. The query client was built using Grails and Javascript, with D3-based graphics (http://d3js.org/), and communicates with the data warehouse via the API. This tool supports predictive text searching, and returns a full set of results for those queries from all compounds, assays, and projects stored within the data warehouse (see Supplementary Information).

The code developed for BARD is free and open-source, and is available on GitHub (see Supplementary Information). Some functions of BARD rely on non-open-source components. Rendering of compound structures within the web interface is provided by the free but closed-source JChem API from ChemAxon (http://www.chemaxon.com/products/jchem-base/). The structure editor is from Scilligence (http://www.scilligence.com/web/jsdrawapis.aspx), and requires a license to be used on a



	Experiment ID	5615									
	Experiment Name	Primary HTS Cell	Broad DOS library								
	result detail	percent activity									
	activity threshold	>=85									
	concentration-point number	1									
	unit	%									
	number of points	1									
	number of replicates	2									
	project lead name	Joshua A. Bittker									
Row #	Substance	Replicate #	Parent Row #	reproducibility	PubChem outcome	percent activity	percent activity	comment	screening concentration (molar)		
1	85789213				Active			2013-01-A01-03-01			
2	85789213		1			142					
3	85789213	1	2				127.843		6.25 uM		
4	85789213	2	2				155.907		6.25 uM		
5	85789213		1	0.995							

Figure 2. Loading data into BARD. New screening experiments can be entered into BARD using the web-based data-entry tool, available on the BARD website. The BARD UI guides the user through the process of creating an assay protocol to describe the experimental design. It then creates an upload template file that contains all of the required fields necessary to deposit a completed set of experimental results.

newly-deployed version of BARD. Without this license, only the structure-editing components of BARD are unavailable. The BARD data-entry tool is supported by an Oracle database (http://www.oracle.com/), and so an Oracle license, held by many institutions, is required for installation and operation of a new instance of BARD.

Depositors

New assay data are entered into BARD directly within a web browser. The BARD website includes a data-entry system that guides data submitters in adding new data to BARD such that it is properly structured and can be included in cross-assay queries (Figure 2). The data-entry interface supports the BARD structured vocabulary by checking input data against the vocabulary and business rules of the assay data standard. User-friendly tools such as auto-suggest and an annotation checklist guide scientists through this process to reduce the effort associated with structuring assay descriptions and ensuring consistent results. A video tutorial demonstrating the structure of the data in BARD and the process of entering data is available online (https://github.com/broadinstitute/BARD/wiki/Video-Demonstrations).

Each experiment entered into BARD is assigned an assay definition, a description of the experiment that includes information such as the assay format (cell-based, biochemical, etc.) and the biological target. When a new assay definition is created, the UI creates basic context items com-

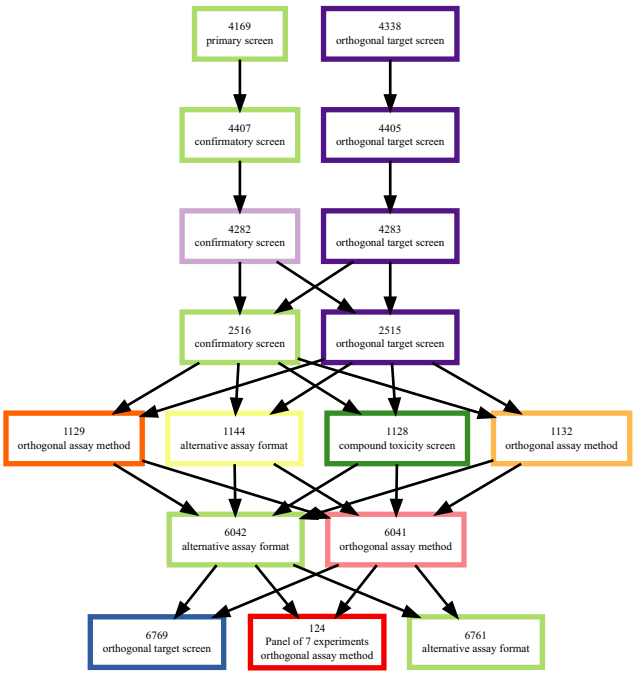


Figure 3. The Project Flow Diagram. The Project Flow Diagram illustrates the relationships between experiments in a project. Experiments like primary screens can be connected to downstream experiments, like confirmatory assays and counter-screens.

mon to all assay definitions. These context items can be entered with further guidance from the UI in the selection of appropriate values for those items: BARD will attempt to auto-complete text entered into the fields with appropriate BARD vocabulary terms. When existing vocabulary terms are not sufficient, new terms can be added. If an identifier from an external database is required (e.g., Uniprot or Entrez Gene ID) the user interface will also ensure valid identifiers are entered. Once completed, an assay definition can be stored and re-used for other experiments.

Once an assay definition is created or chosen from a list of existing assay definitions, BARD will generate a custom tab-delimited text file template that can be used to upload assay result data. Completed experiments can then be linked to other experiments in the project to create a project flow diagram, summarizing the experimental workflow employed by the project team (Figure 3). Properly formed data are transferred to the data warehouse on a daily basis.

Data miners

The data stored in BARD are accessible to the public via three avenues: the browser-based Web Query Client (including plug-ins), the Desktop Client, and third-party standalone applications built by the developer community. Each of the tools allows researchers to retrieve high-quality structured data from BARD.

The Web Query Client features an initial interface built around a single search bar. This search tool offers autosuggest options for targets, gene names, KEGG pathways, GO terms, and other terms from the controlled vocabulary or referenced external ontologies. Queries are performed against the entire database, and results are returned for assay definitions, compounds, projects, and experiments in separate search tabs. These results can be examined and filtered by a variety of criteria such as protein target, GO biological process term, or assay format. Individual items can be stored for later comparison in the Query Cart, a user session-specific storage tool. Results can then be visualized using the web client's built-in Molecular Spreadsheet (Figure 4a) or by downloading to the BARD Desktop Client (Figure 5). The Query Cart can also be used as a workspace, where compounds, assay definitions, and projects can be added and removed as the user progressively refines their search criteria.

Compound search results summarize the number of hits each component has had in BARD-curated assays. Compounds, projects, or assay definitions can be selected and added to the Query Cart for further analysis. In the Molecular Spreadsheet, the detailed results for compounds in various assay definitions can be viewed in a format similar to traditional structure–activity relationship (SAR) tables: a matrix format of compounds versus biology (assay definitions and experiments).

The utility of data curated with well-structured controlled vocabularies is evident in the Linked Hierarchy visualization component of the Web Query Client tool (Figure 4b). The Linked Hierarchy is a set of sunburst or pie charts displaying annotations that describe the current selected compound. The first chart shows the distribution of biological process terms associated with the assays in which the compound was tested. The second chart shows the format of those assays, the third the protein class of the presumed target of those assays, and the fourth the assay type. These four charts are linked; clicking on a pie slice in one chart selects the assays associated with that annotation, and the remaining three charts change layout to display the annotation composition of those selected assays as well. Clicking on the drill-down icon below one of the charts opens up the sunburst display that visualizes the data as a hierarchy (Figure 4b). These trees are strict hierarchies, with each child item having only one parent. Each item in the sunburst representation of these trees is selectable, allowing the user to select a set of terms based on their belonging to a child category of some given parent. By selecting various criteria in each of the pie and sunburst charts, one can make associations between the provided annotations. A video demonstration of the use of the Linked Hierarchy and the Molecular Spreadsheet to generate hypotheses is available online (https://github.com/broadinstitute/BARD/wiki/Video-Demonstrations).

Pre-publication or proprietary data can be compared to public BARD data using the BARD Desktop Client. The tool matches the user's compound structures to those present in the BARD data warehouse by making a hash key of the structure; that hash is compared to the hash keys of BARD compounds so no private structure data need be passed over the network. The Desktop Client is then able to download and present BARD data alongside private data (Figure 5). Since the Desktop Client is run locally, it is not limited by the browser and can load more data than a browser-based tool. It can save queries for later use and will automatically update all downloaded information, en-

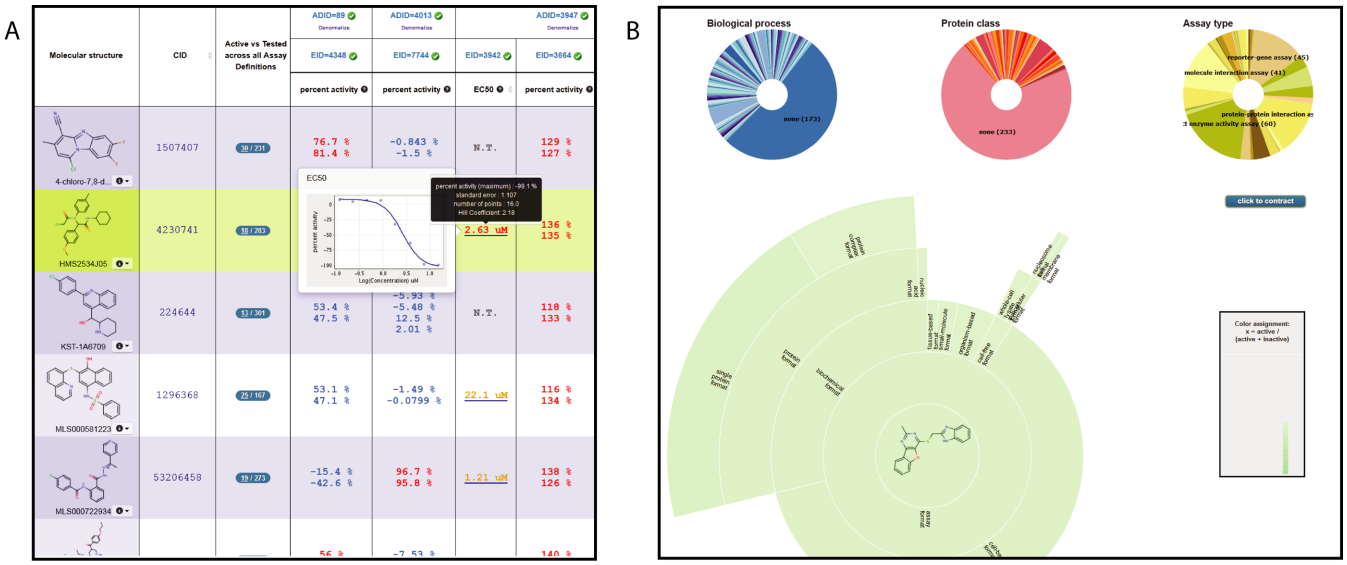


Figure 4. The Web Query Client results. Rich search results are available in the Web Query Client. (**a**) The Molecular Spreadsheet displays a summary table of characteristics of selected compounds, including structure, active versus inactive counts, a complete record of the assays and experiments in which the compounds were active, the concentrations at which they were active, and dose–response curves when they are available. (**b**) The Linked Hierarchy enables investigation of compound annotations. The individual pie or sunburst charts summarize the number of compound hits from assays broken down by several different criteria: biological process, assay format, protein target class, and assay type. These charts are interactive: clicking on an item in the sunburst causes the sunburst to be redrawn to show only those items belonging to the selected item and its children. The remaining three pie charts are also redrawn to display the distribution of items that belong to the selected assay results.

suring that the data are always up to date. Furthermore, the user can export data held in the Desktop Client to third-party applications using the industry-standard structure-definition file (SDF) format.

Developers

The BARD Web Query Client and Desktop Client, while offering functionality to the majority of potential users, are unlikely to support all of the possible needs of all researchers. The BARD API provides two paths for developers to add functionality to BARD: by creating stand-alone tools, and by creating plug-ins that can be deployed as part of the BARD website. The API provides eleven top-level resources that correspond to all the main entities that represent BARD datasets (projects, assays, experiments, etc.). It has complete access to the underlying BARD database, allowing queries of assay data for selected compounds and targets, as well as SAR analyses. Finally, the API is extensible, in that user-contributed code, termed plug-ins, can be deployed within the BARD architecture and presented as a seamless part of the API hierarchy. Since the plug-ins are deployed within the BARD infrastructure, they have direct access to the BARD data warehouse. Links to the full documentation for plug-in development are documented as Supplementary Information.

BARD's plug-in architecture supports a community of researchers that develop new functionality that can be integrated into the BARD website. At least six plug-ins have already been written, including BADAPPLE (Figure 6) (13), a compound promiscuity calculator, and SmartCyp, a tool for predicting compound metabolism by cytochrome P450 (14). This plug-in system creates, in effect, a computational community platform, allowing expansion of the BARD ecosystem of tools and encouraging community involvement in the mining and sharing of bioassay data (see Supplementary Information).

DISCUSSION

BARD addresses a need for a curated, well-structured database to house the wealth of bioassay data produced by the enormous MLP effort in a form that allows effective cross-assay query and analysis. BARD curators have collectively annotated and uploaded over 600 projects, resulting in a data resource unique in its size, quality, and utility. We have improved on existing resources for managing and querying these data by properly annotating and restricting the data inputs to a controlled, yet extensible, vocabulary. The migration process drove the creation of the data-entry component of the BARD web application, resulting in a tool that simplifies the deposit of new data and results in high-quality, well-structured datasets. To support the use of these data, we built a series of data-mining and data-visualization tools as well as an API for programmatic access. These new resources serve three distinct segments of the chemical biology research community: assay data depositors, data miners, and software developers.

The user interface supporting data deposition simplifies the important task of publishing and integrating experimental results and metadata. Data depositors can contribute their experimental data to the BARD database and benefit both themselves and the research community. Upon deposition into BARD, new data immediately become more useful, not just to the research community that gains access to those data, but to the depositor. Because of the consistent structure and restricted vocabulary used in BARD, new datasets are immediately made more powerful by their

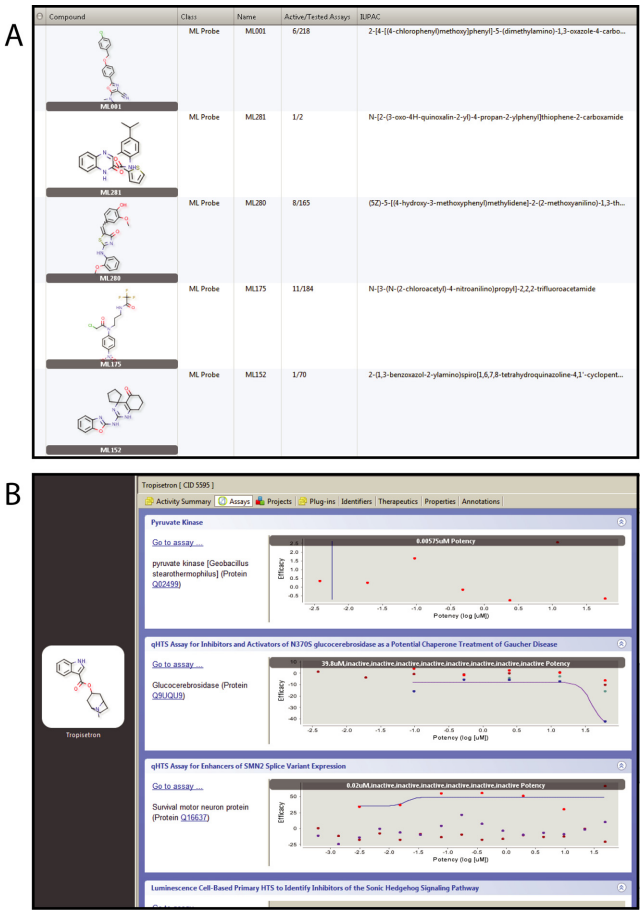


Figure 5. The Desktop Client. The Desktop Client downloads data stored in the user's Query Cart, and displays it alongside local, private data for further analysis. (**a**) The compound-level information is displayed in a tabular format similar to Web Query Client results and includes compound structure and promiscuity information as well as compound class and full name. (**b**) The Assays tab summarizes results for the selected compound in each assay, displaying dose–response curves and other summary statistics.

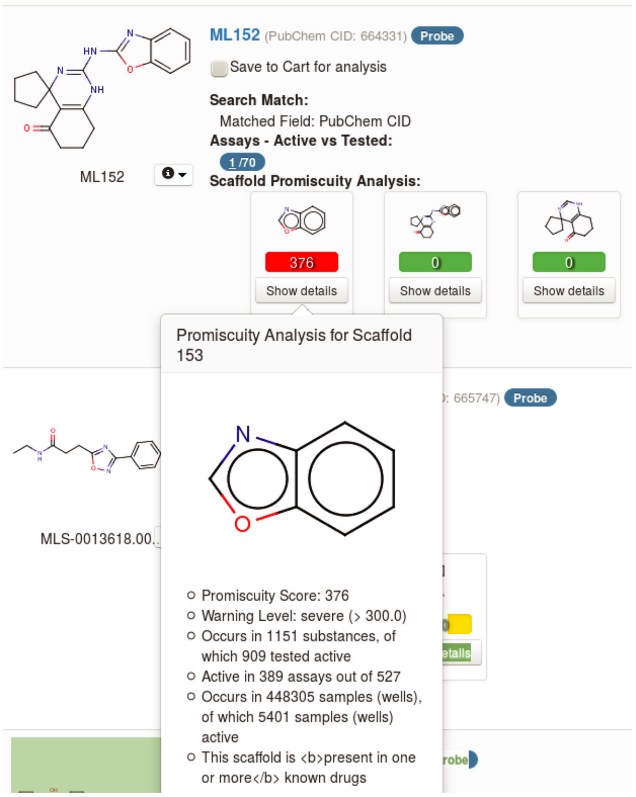


Figure 6. The BADAPPLE plug-in is embedded within the BARD website. The BADAPPLE plug-in accesses the BARD data warehouse using the REST API and implements the plug-in interface, allowing it to be deployed as part of the BARD website. The results are then displayed as part of the compound details. BADAPPLE predicts compound scaffold promiscuity using assay information about compounds containing that scaffold.

association with the existing MLP reference dataset. Such data are now useful to the public and more useful to the researcher who deposited them.

The existing data within BARD is also improved by new data depositions: only searches need be re-run and will automatically include results from newly added data. Furthermore, the data within BARD are improved when updates to established data resources are made available. For instance, when the GO database is updated, it is automatically integrated into BARD and associations of both old and new bioassay data are made with the updated ontology.

The extensive data-access and data-analysis tools within the Web Query Client, Desktop Client, and the various plug-ins support researchers in mining the data and generating hypotheses. With these tools, researchers can learn about themes in compound activity (e.g., investigating enrichment of kinase-inhibitory compounds for various target biological processes). The Linked Hierarchy promotes concise, rapid search, sort, and visualization of these data in a way that allows filtering of multiple fields, allowing hypotheses to be generated and refined rapidly. The Desktop Client allows the investigation of private or very large datasets alongside publicly available ones.

Finally, the data contained within BARD's structured warehouse can be harnessed by new tools developed by the community—the larger population of researchers and developers—through the use of the API. New tools, either stand-alone or deployed as part of the BARD website, can leverage the structure of the BARD data and expand the analyses available to the greater research community. The extensible plug-in architecture allows a new community of tool builders to form around the data.

We built BARD with an understanding of the importance of data sharing in the screening and chemical biology research community, and so we designed it to ease the process of data-deposition with a high-quality data-entry toolkit. Further, we recognize that depositing data into a public repository is often not a high priority for researchers, so we built a system that has value to the data depositor themselves, increasing the power of a researcher's own data because of the context provided by the larger BARD dataset. We hope that these tools will encourage the submission of many future datasets to BARD. In summary, we created a powerful query engine for hypothesis generation and probe development, including data and access tools that can make probe discovery faster, cheaper and easier, leading to improvements in human health.

ACKNOWLEDGMENTS

The authors would like to thank Dr Christopher Austin for his commitment to the preservation and reuse of research data, and his continuing support of BARD in particular, which directly benefited from his expertise and leadership.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Institutes of Health Common Fund through its Molecular Libraries Probe Production Center Network [RFA-RM-08-005 to BARD] as an administrative supplement directly to the centers following receipt and review of center proposals in response to a request for proposal crafted by Dr. Ajay Pillai (National Human Genome Research Institute). Funding for open access charge: Broad Institute Comprehensive Screening Center (U54-HG005032).

Conflict of interest statement. E.A.H., A.D., D.L.L., X.X., J.A., D.D. and J.A.B. are considering a commercial data sciences enterprise that would partner with academic and bio-medical research institutes to commercialize software. This enterprise would use elements of the BARD toolkit and migrated data along with elements of other open source projects.

REFERENCES

1. Strausberg,R.L. and Schreiber,S.L. (2003) From knowing to controlling: a path from genomics to drugs using small molecule probes. *Science*, **300**, 294–295.
2. Seiler,K.P., George,G.A., Happ,M.P., Bodycombe,N.E., Carrinski,H.A., Norton,S., Brudz,S., Sullivan,J.P., Muhlich,J., Serrano,M. *et al.* (2008) ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.*, **36**, D351–D359.
3. Bolton,E.E., Wang,Y., Thiessen,P.A. and Bryant,S.H. (2008) *PubChem: Integrated Platform of Small Molecules and Biological Activities*. Elsevier, **4**, pp. 217–241.
4. Wang,Y., Xiao,J., Suzek,T.O., Zhang,J., Wang,J. and Bryant,S.H. (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.*, **37**, W623–W633.
5. Wang,Y., Xiao,J., Suzek,T.O., Zhang,J., Wang,J., Zhou,Z., Han,L., Karapetyan,K., Dracheva,S., Shoemaker,B.A. *et al.* (2012) PubChem's BioAssay Database. *Nucleic Acids Res.*, **40**, D400–D412.
6. Bento,A.P., Gaulton,A., Hersey,A., Bellis,L.J., Chambers,J., Davies,M., Krüger,F.A., Light,Y., Mak,L., McGlinchey,S. *et al.* (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Res.*, **42**, D1083–D1090.
7. Gaulton,A., Bellis,L.J., Bento,A.P., Chambers,J., Davies,M., Hersey,A., Light,Y., McGlinchey,S., Michalovich,D., Al-Lazikani,B. *et al.* (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, D1100–D1107.
8. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
9. Kanehisa,M., Goto,S., Sato,Y., Furumichi,M. and Tanabe,M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.
10. de Souza,A., Bittker,J.A., Lahr,D.L., Brudz,S., Chatwin,S., Oprea,T.I., Waller,A., Yang,J.J., Southall,N., Guha,R. *et al.* (2014) An Overview of the Challenges in Designing, Integrating, and Delivering BARD: A Public Chemical-Biology Resource and Query Portal for Multiple Organizations, Locations, and Disciplines. *J. Biomol. Screen.*, **19**, 614–627.
11. Visser,U., Abeyruwan,S., Vempati,U., Smith,R.P., Lemmon,V. and Schürer,S.C. (2011) BioAssay Ontology (BAO): a semantic description of bioassays and high-throughput screening results. *BMC Bioinformatics*, **12**, 257.
12. Vempati,U.D., Przydzial,M.J., Chung,C., Abeyruwan,S., Mir,A., Sakurai,K., Visser,U., Lemmon,V.P. and Schürer,S.C. (2012) Formalization, annotation and analysis of diverse drug and probe screening assay datasets using the BioAssay Ontology (BAO). *PLoS ONE*, **7**, 49198.
13. Yang,J. (2013) The badapple promiscuity plugin for bard: Evidence-based promiscuity scores. In: *ACS National Meeting*.
14. Rydberg,P., Gloriam,D.E., Zaretzki,J., Breneman,C. and Olsen,L. (2010) SMARTCyp: A 2D Method for Prediction of Cytochrome P450-Mediated Drug Metabolism. *ACS Med. Chem. Lett.*, **1**, 96–100.